PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2006) **273**, 401–407
doi:10.1098/rspb.2005.3263
Published online 15 November 2005

An unusual choanoflagellate protein released by Hedgehog autocatalytic processing

Elizabeth A. Snell[1,2], Nina M. Brooke[1,3,4], William R. Taylor[5], Didier Casane[6,7], Hervé Philippe[6,8] and Peter W. H. Holland[3,*]

[1]*School of Animal and Microbial Sciences, The University of Reading, Whiteknights, Reading RG6 6AJ, UK*
[2]*Department of Biology, North Carolina A&T State University, Greensboro, NC 27411, USA*
[3]*Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK*
[4]*School of Continuing Education, The University of Reading, London Road, Reading RG1 5AQ, UK*
[5]*The National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK*
[6]*Phylogénie, Bioinformatique et Génome, UMR 7622 CNRS, Université Pierre et Marie Curie, 9 quai St Bernard Bât. C, 75005 Paris, France*
[7]*Populations, Genetique et Evolution, UPR9034, CNRS, 91198 Gif-sur-Yvette, France*
[8]*Département de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C3J7, Canada*

Hedgehog proteins are important cell–cell signalling proteins utilized during the development of multicellular animals. Members of the *hedgehog* gene family have not been detected outside the Metazoa, raising unanswered questions about their evolutionary origin. Here we report a highly unusual *hedgehog*-related gene from a choanoflagellate, a close unicellular relative of the animals. The deduced C-terminal domain, Hoglet-C, is homologous to the autocatalytic domain of Hedgehog proteins and is predicted to function in autocatalytic cleavage of the precursor peptide. In contrast, the N-terminal Hoglet-N peptide has no similarity to the signalling peptide of Hedgehog (Hh-N). Instead, Hoglet-N is deduced to be a secreted protein with an enormous threonine-rich domain of unprecedented size and purity (over 200 threonine residues) and two polysaccharide-binding domains. Structural modelling reveals that these domains have a novel combination of features found in cellulose-binding domains (CBD) of types IIa and IIb, and are expected to bind cellulose. We propose that the two CBD domains enable Hoglet-N to bind to plant matter, tethering an amorphous nucleophilic anchor, facilitating transient adhesion of the choanoflagellate cell. Since Hh-C and Hoglet-C are homologous, but Hh-N and Hoglet-N are not, we argue that metazoan *hedgehog* genes evolved by fusion of two distinct genes.

Keywords: *Monosiga*; multicellularity; adhesion; hoglet; cellulose-binding domains

1. INTRODUCTION

Multicellular animals, or metazoans, use many distinct signalling proteins for cell-to-cell communication. These are encoded by a small number of gene families, notably the *hedgehog*, *BMP*, *Wnt*, *FGF*, *TGFβ* and *ephrin* families. To understand how multicellular animals evolved from their unicellular ancestors, it is necessary to resolve the ancestry of each of the gene families encoding signalling proteins.

The *hedgehog* gene family is a particular enigma. This is a small gene family, with single genes reported from *Drosophila* and amphioxus, two genes in *Ciona* and three or more genes in vertebrates (Lee *et al.* 1992; Zardoya *et al.* 1996; Shimeld 1999; Hino *et al.* 2003). It is sometimes argued that there is no *hedgehog* gene in the genome of the nematode *Caenorhabditis elegans*, but in fact there are many related genes (the *warthog* and *groundhog* genes) that probably evolved by duplication and divergence from an ancestral *hedgehog* gene (Aspöck *et al.* 1999). Hedgehog proteins are unusual in undergoing autocatalytic cleavage, yielding a secreted N-terminal signalling domain (Hh-N). The C-terminal domain (Hh-C) is responsible for catalysing the protein cleavage reaction (Porter *et al.* 1995; Hall *et al.* 1997). No *hedgehog* genes have been cloned from outside the animal kingdom, so the evolutionary origins of this gene family are obscure. One clue comes from weak sequence and structural similarity between part of Hh-C and the inteins of bacteria, algal organelles and fungi: a group of proteins that catalyse their own excision from precursor proteins (Hall *et al.* 1997). The region of sequence similarity is called the Hint domain (Hedgehog, intein). On the basis of this structural similarity, it has been proposed that the *hedgehog* gene family and the inteins evolved from a common precursor.

Choanoflagellates are unicellular protists that are close relatives of multicellular animals (James-Clark 1866; King & Carroll 2001; Snell *et al.* 2001; Lang *et al.* 2002; Philippe *et al.* 2004). Comparison between choanoflagellates and metazoans should allow insight into the origins of animal genes and animal development. Here we describe an unusual gene from a freshwater choanoflagellate *Monosiga ovata*. Sequence of this gene, *hoglet*, implies it uses the Hedgehog autoprocessing mechanism to release a secreted protein with a fundamentally different role from metazoan Hedgehog proteins.

* Author for correspondence (peter.holland@zoo.ox.ac.uk).

Received 20 May 2005
Accepted 24 July 2005

2. MATERIAL AND METHODS

(a) *Gene cloning*

Monosiga ovata were obtained from the American Type Culture Collection (www.atcc.org; ATCC strain number 50635). As part of a phylogenetic study, we performed DNA sequencing on 1600 clones from an amplified cDNA library (Philippe *et al.* 2004). Three sequences had significant basic local alignment search tool (BLAST) matches to the *hedgehog* gene family. To extend this sequence, cDNA library screening was used, followed by PCR from library template and 5′ RACE System v. 2.0 (Gibco BRL). Library PCR used the SP6 vector primer, with an internal primer Nhh4R (5′-ATGCTAACG-GAAGAATCCCA-3′). For rapid amplification of cDNA ends (RACE), four gene-specific primers were designed close to the 5′ end of the longest cDNA clone: hhGSP1 (5′-GTGACAG-TAGCGTGGTCACTGGAATAG-3′), hhGSP2 (5′-CTGA AATCCCACTCAAACTGGAAACCT-3′), hhGSP3 (5′-AC CACAGGGGGTGCCA GCAATGGAAAAT-3′) and Nhh 5R (5′-CAGTAGCGT GGTCACTGGAA). Primer hhGSP3 was used to prime cDNA synthesis, before linker addition and PCR amplification using hhGSP1 with the supplier's AUAP primer. In addition, Nhh5R was used for cDNA priming, followed by PCR using hhGSP2 and AUAP. Products of all PCR amplifications were cloned into pGEM T-Easy vector (Promega), and multiple clones of each type sequenced using an ABI 3100 sequencer. To map intron positions, genomic DNA was extracted from a 100 ml culture of *M. ovata* using Tri-Reagent (Sigma) and used as a template for PCR amplification. Amplified products were cloned and sequenced as above. Sequences reported in this paper have been deposited on GenBank under accession numbers DQ191761–3.

(b) *Sequence analysis*

For phylogenetic analysis of Hoglet-N, Hh-N and inteins, translated protein sequences were aligned using ClustalX, the intein endonuclease region removed and regions of ambiguous homology removed, creating an edited alignment of 51 sites for 16 sequences. For tree reconstruction, we first applied ProtTest (Abascal *et al.* 2005) to estimate the optimal model of amino acid substitution (consistently found to be WAG+I+G), then calculated a maximum likelihood tree using PHYML (Guindon & Gascuel 2003) with this model (the proportion of invariant sites calculated from the alignment, and four rate categories with a gamma distribution parameter estimated from the data). Signal sequences were predicted using SignalP v. 3.0 (http://www.cbs.dtu.dk/services/SignalP-3.0/; Bendtsen *et al.* 2004) and SigCleave (http://bioweb.pasteur.fr/seqanal/interfaces/sigcleave.html). Homopolymeric runs were accessed at the TRIPS database (http://www.ncl-india.org/trips/index.html; Katti *et al.* 2000), compiled from SwissProt Release 38, followed by checking updated sequences in SwissProt Release 44. SwissProt Release 44 (July 2004) was also searched specifically for polyThr repeats by Dr Mukund Katti.

(c) *Structural modelling*

Modelling of the two amino-terminal cellulose-binding domains (CBDs) followed the protocol used by Taylor & Stoye (2004), employing the multiple-sequence/structure alignment (threading) program MST (Taylor 1997). This approach matches a multiple sequence alignment containing a protein of unknown structure onto a known protein structure, simultaneously optimizing the match of predicted and observed secondary structure, hydrophobic burial, residue packing and sequence similarity. To identify potential template structures, the sequences of the two predicted CBD domains were independently scanned against the Protein Data-Bank (PDB) using the program GenTHREADER (Jones 1999). Models were depicted using RasMol.

3. RESULTS

(a) *Organization of the* hoglet *gene*

In the course of an expressed sequence tag (EST) sequencing project on the unicellular choanoflagellate *M. ovata* (Philippe *et al.* 2004), we identified three partial cDNA clones with significant BLAST similarity matches to the *hedgehog* gene family (*M. ovata* clones 5G4, 5H8, 4H6); this gene family includes the *hedgehog* gene of *Drosophila* and the vertebrate *Sonic hedgehog*, *Indian hedgehog* and *Desert hedgehog* genes. The region of similarity was located exclusively in the C-terminal region of Hedgehog proteins (the Hint domain, comprising a large part of Hh-C). In Hedgehog proteins, this domain catalyses cleavage of the protein between glycine and cysteine residues of a conserved Gly-Cys-Phe (GCF) motif. The *Monosiga* gene possesses the GCF motif, as well as two absolutely conserved residues shown to be required for thioester formation and intramolecular cleavage of Hedgehog proteins: Thr326 and His329 (Hall *et al.* 1997). This strongly suggests that the choanoflagellate gene also encodes a protein capable of autocatalytic cleavage. We name this choanoflagellate gene *hoglet*.

The initial *Monosiga* cDNA clones were truncated at their 5′ end. Screening of the cDNA library by hybridization yielded six longer partial cDNA clones, of which only one extended beyond a huge $(ACN)_n$ repeat. We then used PCR versus cDNA library template, followed by 5′ RACE PCR, to obtain the full-length cDNA sequence. This comprises a short 5′ untranslated region (UTR) of 15 nucleotides, an in-frame methionine with perfect Kozak sequence (CCACCATGG), an open reading frame of 1989 nucleotides and 3′ UTR of 61 nucleotides (figure 1).

Fifteen nucleotides is generally thought to be the minimum possible 5′ UTR length, because any shorter would not allow the P-site of the ribosomal 43S subunit to recognize the AUG while also interacting with the cap-binding protein complex. Nonetheless, even this length restriction can be overcome in exceptional cases; *Giardia* (a diplomonad) has many 5′ UTRs less than 15 nucleotides in length, even down to zero nucleotides (Adam 2000). Using PCR on genomic DNA, we mapped two introns in the *hoglet* gene, of 146 and 142 nucleotides length (figure 1).

The *hoglet* gene is predicted to encode a protein product of 663 amino acids, which will undergo autocatalytic cleavage at a conserved GCF motif to yield an N-terminal peptide of 450 amino acids (Hoglet-N) and a C-terminal peptide of 213 amino acids (Hoglet-C); figure 2. Each of these peptides has surprising features.

(b) *Hoglet-C: homology to Hh-C*

Hoglet-C is clearly similar to the Hh-C domain of Hedgehog proteins. It is not an intein, because the Hint domain is close to the C-terminus, compatible with a single site of autocatalytic cleavage, and because it lacks

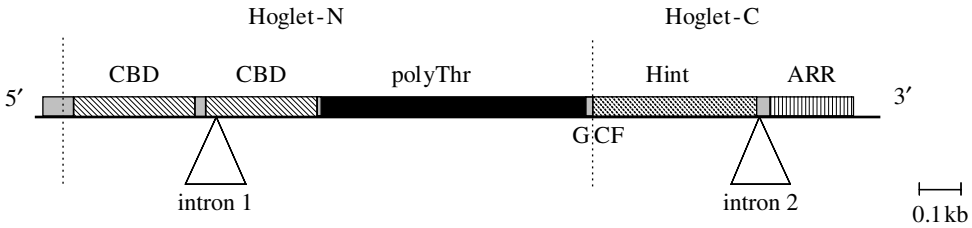

Figure 1. Organization of *Monosiga ovata hoglet*. The coding sequence is shown as a rectangle; within this, CDB domains have diagonal hatching, polyThr repeat is shaded black, Hint domain is speckled and the adduct recognition region (ARR) has vertical hatching. The intersect vertical lines denote two predicted protein cleavage sites; triangles denote introns.

```
  1  MAPCGIVAALLVLSASAVLATPPACSDVTYTQTARWQGGFVGTFTINNAGTGFQFEWDFS
 61  DFGYPVTLASGDVWNANLDSYSSDHATVTLDADQNSFGFVAGTEAPPGAVRAADVGGFSI
121  AGTPCGGVATTPAPTACSNLNYQVTASWEDGFVIRVDLQGGAQNSIIEWTWDSSVSIQGS
181  VWGADVLSSSGNTVKFGVAGESFGMLLSGSSTTINSIQINGNPCTGGGSTTTSTSSTTTT
241  TTTTSSTTSTTTTTTTSTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGTTTTTTTGSTTTTT
301  TTTGSTTTTTTTTTTTSTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
361  TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
421  TTTTTTTTTTTTTTTTTTTTTTTTTPTTTFGCFSQSSTVVVEGRGRISITDVQPGDMLVDG
481  SAAGTSQFVSFMHVGEQLAPFVAIQAGNNTIKLTALHLLYVVDAKQGAELVKASEVRVGN
541  TVILADGEHAIVDAVSIVEEEGFISVLTASGQLAVDGIVASSYTASKVYPGEPISLRHAL
601  NQPLIWLAQAAPSLAAPMLSSEYHWYQQVTRAPRAMAHAFRSIAKAVEARMPALATVSCT
661  GSL
```

Figure 2. Complete deduced protein sequence of *Monosiga ovata* Hoglet. Bold italics denote the signal sequence and the GCF motif. The stars denote the two predicted protein cleavage sites. Triangles denote intron positions (the first between Q and V codons, second between K and V codons). The CDB domains are underlined; the extensive polyThr repeat has dotted underlining. DNA sequences have been deposited on GenBank/EMBL/DDBJ under accession numbers DQ191761 (cDNA) and DQ191762-3 (introns).

a core endonuclease or linker region within the Hint module. Molecular phylogenetic analysis also reveals a closer evolutionary relationship between *hoglet* and *hedgehog* genes than between *hoglet* and inteins (figure 3).

Autocatalytic cleavage of Hedgehog proteins requires the involvement of cholesterol during cleavage, followed by covalent linkage between cholesterol and the Hh-N peptide. The binding of cholesterol is effected by a sterol recognition region (SRR) located immediately C-terminal to the Hint domain. This domain is not highly conserved; notably, the equivalent domain in nematode Groundhog and Warthog proteins shares principally the clustering and spacing of hydrophobic residues, and has been designated the adduct recognition region (ARR; Mann & Beachy 2000). Hoglet-C has a stretch of similar composition in the equivalent position, possibly homologous to the ARR and SRR, although it is questionable whether sequence similarity is sufficient for effecting the same function (figures 1 and 2).

(c) *Hoglet-N: polysaccharide-binding domains and polyThr repeat*

Analysis of the Hoglet-N sequence indicates that the first 20 amino acids act as a cleaved signal sequence (Signal P probability 1.000; Sigcleave minweight score 9.6, scores greater than 3.5 indicate signal peptides). The deduced 430 amino acid secreted peptide has no sequence similarity to the secreted portion of metazoan Hedgehog proteins. The peptide has some remarkable sequence characteristics. Included within the 430 residues is a 219 amino acid stretch containing an astonishing 205 threonines, including an unbroken homopolymeric run of 128 threonine residues (figure 2). Although repeated runs of amino acids are commonly found in proteins, the length and purity of this polyThr repeat is completely without precedent. The longest homopolymer runs described by Katti *et al.* (2000), in a catalogue of tandem repeats in proteins, are a run of 52/54 asparagines (in *Dictyostelium* Protein Tyrosine Phosphatase 3) and 51/55 glutamines (in yeast SNF5). PolyThr repeats are typically much shorter, the maximum identified previously being 25/28 (M. Katti 2004, personal communication). None of these repeats come close in length to the 205/219 threonines present in Hoglet-N.

N-terminal to the polyThr repeat are two tandemly arranged motifs that are clearly recognizable as family II CBDs (Tomme *et al.* 1995) delineated by Cys residues (figure 2). Each domain has three Trp residues at positions

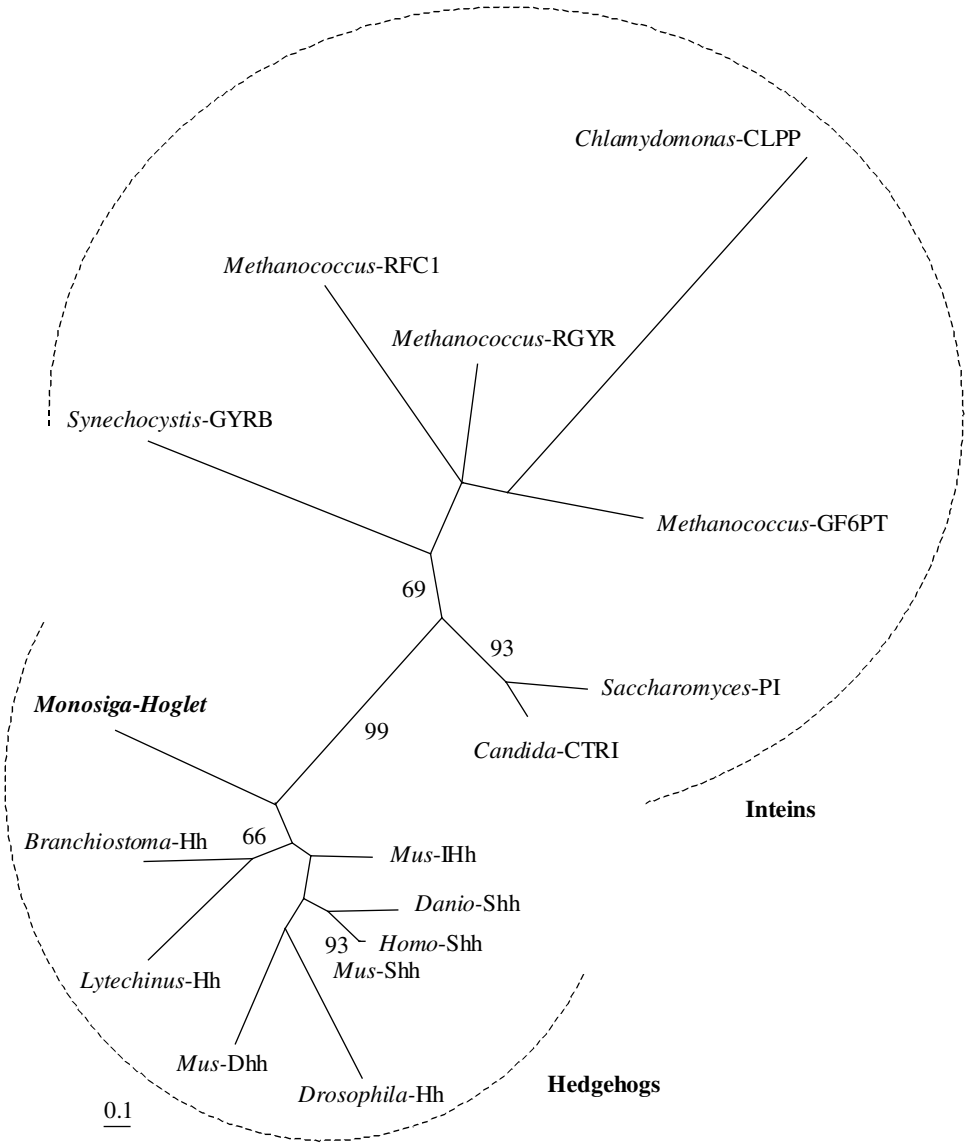

Figure 3. Unrooted phylogenetic tree showing the evolutionary relationship between Hoglet-C, Hedgehog proteins and inteins. Resolution within the *hedgehog* gene family, and between *hedgehogs* and *hoglet*, is compromised by the limited alignment possible to inteins; nonetheless, Hoglet-C is more closely related to Hedgehog-C than to inteins. Numbers denote support values from 100 bootstrap resamplings of the data. Abbreviations of proteins: Hh, Hedgehog; Shh, Sonic Hedgehog; Ihh, Indian Hedgehog; Dhh, Desert Hedgehog; GF6PT, glutamine fructose 6-phosphate transaminase; RFC1, replication factor C; RYGR, reverse gyrase; GYRB, DNA gyrase subunit B; CLPP, ClpP protease. Species: *Monosiga ovata*; *Homo sapiens*; *Mus musculus*; *Danio rerio*; *Branchiostoma floridae*; *Lytechinus variegatus* (sea urchin); *Drosophila melanogaster*; *Saccharomyces cerevisiae*; *Candida tropicalis* (yeast); *Chlamydomonas eugametos* (green alga); *Synechocystis* sp. PCC 6803 (Eubacteria); *Methanococcus jannaschii* (Archaea). CTRI and PI, intein/endonuclease of vacuolar H^+-ATPase.

conserved with other members of the CBD II family: Trp36, Trp57 and Trp74 in domain 1, and Trp148, Trp169 and Trp182 in domain 2. Exposed Trp residues play a direct role in binding of CBD II domains to polysaccharide. Tomme *et al.* (1995) divide the CBD II family into CBD IIa and CBD IIb, the latter having a C-terminal deletion covering an otherwise conserved Trp residue. On this criterion, the CBD domains of Hoglet-N are in the CBD IIb subfamily (also called XBD, xylan-binding domains, reflecting binding specificity of some domains; Dupont *et al.* 1998).

(**d**) *Structural modelling of CDB domains*

To predict the structure of the two CBD domains from Hoglet-N, and inform probable biological role, we scanned each domain against the Protein Data Bank using GenTHREADER (Jones 1999). Each gave only two matches in the high or medium confidence range; these were CBD IIa (PDB identifier 1exg) and XBD (PDB identifier 1xbd). For domain 1, 1exg gave a score of 0.728 (high); 1xbd gave a score of 0.685 (high). For domain 2, 1exg gave a score of 0.633 (medium); 1xbd gave a score of 0.621 (medium). Consequently, models were constructed for domain 1 and domain 2, using both 1exg and 1xbd as templates. Comparing the modelled structures showed that choice of template affected the models, particularly in the conformation of the C-terminal beta strand (for domain 1, 5.5 Å root mean squared difference over 102 residues; for domain 2, 4.1 Å over 94 residues). Examination of the primary sequence strongly suggests this C-terminal strand will fold as in CBD IIa domains, because of a conserved #x#xGxPC

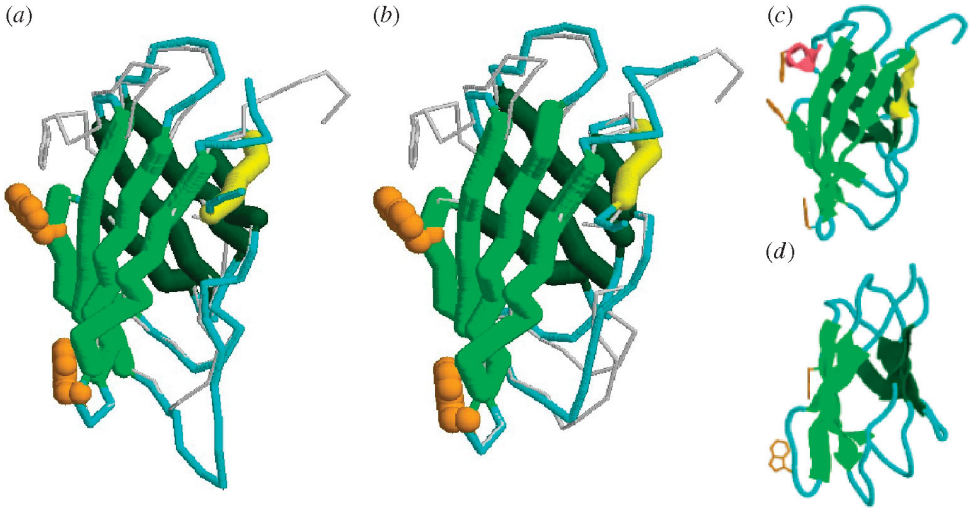

Figure 4. (*a*) The structure of the first Hoglet polysaccharide-binding domain modelled on a CDB IIa domain (PDB code 1exg, shown in grey). The two beta sheets are coloured light and dark green, while a C-terminus beta-strand linking the two sheets is yellow. The side chains of the two exposed tryptophans are shown in orange. Loops are shown in blue. (*b*) The structure of the second Hoglet polysaccharide-binding domain modelled on 1exg, coloured as in (*a*). This shorter domain has essentially the same structure, but with additional deletions in some loops. (*c*) The structure of the cellulose-binding domain of exo-1,4-beta-D-glycanase from *Cellulomonas fimi* (PDB code 1exg), used as a template for the Hoglet domain structures. This structure differs from both Hoglet domains principally by presence of a mini alpha-helix (pink), including another exposed tryptophan. However, like the Hoglet domains the lower two tryptophans, implicated in polysaccharide binding, have coplanar orientation suitable for cellulose binding. (*d*) The structure of the xylan-binding domain (CBD IIb) of the endo-1,4-beta-D-glycanase from *C. fimi* (PDB code: 1xbd). Like the Hoglet domains, the mini-helix is not present. However, the two tryptophans have an orthogonal orientation suited to binding xylan.

motif (#=hydrophobic) in Hoglet-N domains and CBD IIa proteins, in which the hydrophobic domains are buried in the 1exg structure. Four loops between beta strands differ in length between the 1exg and 1xbd structures; in all but one case the 1exg structure proved a more suitable template for modelling the Hoglet domain sequences. The exception is the loop including a solvent-exposed Trp that is deleted in Hoglet and the CBD IIb family, noted above. In this region, the shorter loop in 1xbd provided a better template; however, this deletion has little effect on the overall structure. Therefore, 1exg can be used as a suitable template for modelling both CBD domain 1 and domain 2 of Hoglet-N (figure 4*a*,*b*).

We deduce that each CBD domain in Hoglet-N will fold into two beta sheets forming a twisted beta-sandwich motif, with two solvent-exposed Trp residues (Trp36 and Trp74 in domain 1; Trp148 and Trp182 in domain 2). These exposed Trp residues correspond to Trp259 and Trp291 in the XBD of *Cellulomonas fimi* xylanase D (numbering follows Simpson *et al.* 1999) and to Trp17 and Trp54 of *C. fimi* Cex (numbering follows Bray *et al.* 1996). These residues are involved in binding to xylan and to cellulose, respectively.

An unexpected feature of the Hoglet structure concerns the orientation of the exposed Trp residues. These residues are coplanar in the Hoglet domains (figure 4*a*,*b*) and in the CBD of Cex (1exg structure; figure 4*c*), but are at 90° to each other in the xylan-binding domain of XBD1 (1xbd structure; figure 4*d*). These different orientations are thought to allow binding to either the flat polymer cellulose, or the twisted polymer xylan, through stacking between Trp residues and sugar moieties (Simpson *et al.* 1999). Hence, although the Hoglet CBD domains associate more with the CBD IIb (XBD) family on the basis of primary sequence, they are predicted to bind cellulose not xylan.

4. DISCUSSION

We report a highly unusual *hedgehog*-related gene from a single-celled choanoflagellate. The C-terminal domain of the deduced protein has significant similarity to the autocatalytic domain of Hedgehog and is expected to catalyse cleavage at a conserved GCF site to release a secreted peptide. Nonetheless, we argue that this gene, denoted *hoglet*, should not be placed within the *hedgehog* gene family, because there is no similarity in the N-terminal domain of the protein.

We do not know the precise biological role of secreted Hoglet-N, but its unusual protein sequence allows realistic suggestions to be made. The protein contains two CBDs, followed by a huge polyThr repeat of unprecedented size and purity. Modelling indicates that the two CBD domains have a novel structure with features characteristic of both CBD IIa and CBD IIb subfamilies. Importantly, each domain has two solvent-exposed tryptophan residues with coplanar orientation, as in the CBD IIa domain of *C. fimi* Cex, and unlike the mutually perpendicular orientations in xylanase D. This implies the Hoglet-N domains are suitable for binding to the planar polymer cellulose (found in plant cell walls and some bacteria), rather than the more twisted xylan. Most (but not all) CBD proteins possess an enzymatic domain, enabling cellulose or xylan degradation. Hoglet-N does not have an enzymatic domain, so cannot be directly involved in cellulose degradation. Instead, the biochemical function of the two CBD domains in Hoglet-N must be to bind with high affinity to plant or algal cell walls (abundant in aquatic habitats), tethering the enormous polyThr domain.

The function of the tethered polyThr domain is less clear. Hwang & Stupp (1999) report that aqueous synthetic polythreonine forms a liquid crystalline matrix when in contact with biological tissues. This matrix impedes wetting, which in turn increases the efficacy of

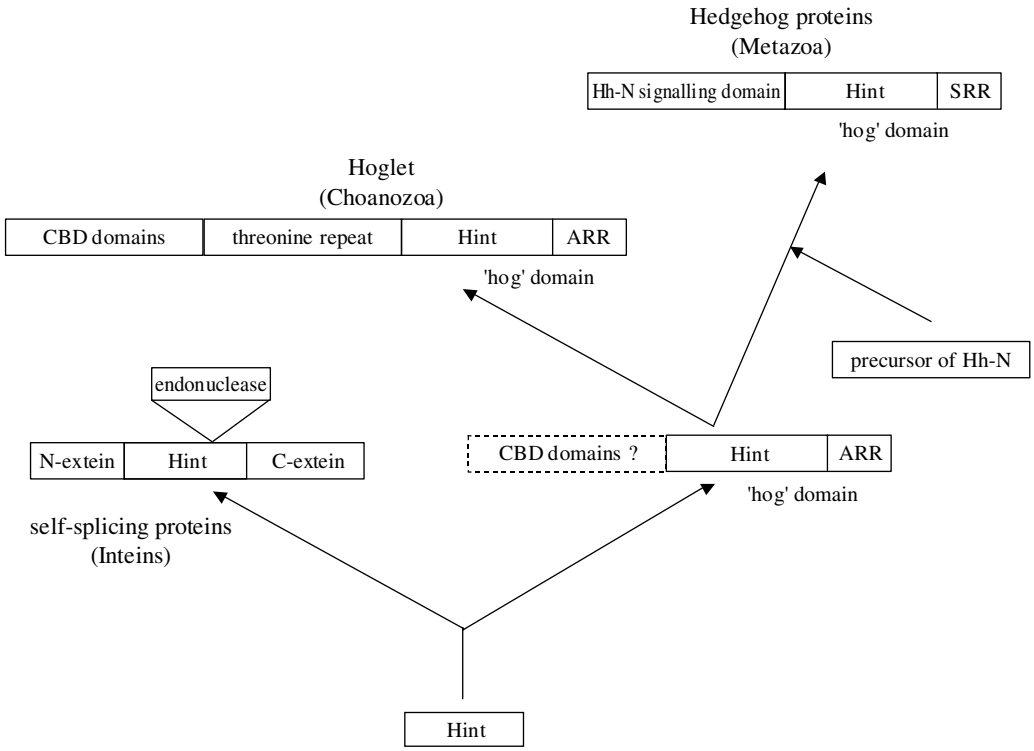

Figure 5. Schematic model for the evolution of metazoan Hedgehog proteins. In the scenario proposed, Hedgehog family genes were assembled from two ancestral genes, the C-terminal derived from an Hint domain-containing gene in the common ancestor of Choanozoa and Metazoa. The N-terminal peptide of this ancestral gene may have encoded CBD domains; the C-terminal contained an adduct recognition region (ARR) and catalysed cleavage.

glues (e.g. in surgical procedures). This is particularly effective with polythreonine of molecular weight between 5 and 50 kD, a range encompassing the size of the polyThr stretch in Hoglet-N (21.9 kD). It is likely that the polyThr repeat folds into an amorphous domain presenting a large surface of hydroxyl groups. This nucleophilic domain would be adhesive, perhaps permitting transient adhesion of the choanoflagellate to plant and algal cells. Consistent with this, solitary choanoflagellates such as *M. ovata* frequently adhere transiently to their substrate, thereby allowing water currents generated by flagellar motion to bring bacterial food particles to the collar of tentacles (Pettitt *et al.* 2002; N. M. Brooke and E. A. Snell 2003, unpublished observations).

Since choanoflagellates are the sister group to the multicellular animals, we can reasonably ask whether *hoglet* gives any insight into the evolutionary origin of the *hedgehog* gene family. The common ancestor of choanoflagellates and animals was certainly a single-celled organism, and possibly similar in morphology and cellular properties to some living choanoflagellates (James-Clark 1866; Philippe *et al.* 2004). This ancestor could conceivably have possessed a *hedgehog* gene, either lost or still present (but undetected) in choanoflagellates; however, this would raise the question of its role in a single-celled organism. It seems more plausible that this extinct organism possessed a *hoglet*-type gene, or a gene ancestral to *hedgehog* and *hoglet*. If the *hedgehog* gene family did evolve from an ancestral *hoglet*-like gene, it is unlikely that the Hh-N domain evolved directly from a CBD domain protein, because the tertiary structures of these domains are very different. It is more likely that true *hedgehog* genes were assembled on the metazoan lineage from components of two different genes (figure 5): the Hh-C domain from a *hoglet*-like precursor; and the Hh-N domain from different gene (possibly a zinc hydrolase; Hall *et al.* 1995). We propose, but cannot be certain, that CBD domains were present in this ancestral protein. The hybrid gene was then co-opted for a novel function, namely cell–cell signalling during multicellular development.

This research was funded by the BBSRC, MRC and HFSP grant RGP221/2001. We thank Dr Mukund Katti (National Chemical Laboratory, Pune, India) for performing searches for polyThr proteins, and Bernd Schierwater and Steve Dellaporta for HFSP collaboration.

REFERENCES

Abascal, F., Zardoya, R. & Posada, D. 2005 ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105. (doi:10.1093/bioinformatics/bti263)

Adam, R. D. 2000 The *Giardia lamblia* genome. *Int. J. Parasitol.* **30**, 475–484. (doi:10.1016/S0020-7519(99)00191-5)

Aspöck, G., Kagoshima, H., Niklaus, G. & Bürglin, T. R. 1999 *Caenorhabditis elegans* has scores of *hedgehog*-related genes: sequence and expression analysis. *Genome Res.* **9**, 909–923. (doi:10.1101/gr.9.10.909)

Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. 2004 Improved prediction of signal peptides—SignalP 3.0. *J. Mol. Biol.* **340**, 783–795. (doi:10.1016/j.jmb.2004.05.028)

Bray, M. R., Johnson, P. E., Gilkes, N. R., McIntsoh, L. P., Kilburn, D. G. & Warren, R. A. J. 1996 Probing the role of trytophan residues in a cellulose-binding domain by chemical modification. *Protein Sci.* **5**, 2311–2318.

Dupont, C., Roberge, M., Shareck, F., Morosoli, R. & Kluepfel, F. 1998 Substrate-binding domains of glycanases from *Streptomyces lividans*: characterization of a new family of xylan-binding domains. *Biochem. J.* **330**, 41–45.

Guindon, S. & Gascuel, O. 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704. (doi:10.1080/10635150390235520)

Hall, T. M. T., Porter, J. A., Beachy, P. A. & Leahy, D. J. 1995 A potential catalytic site revealed by the 1.7-Å crystal structure of the amino-terminal signalling domain of Sonic hedgehog. *Nature* **378**, 212–216. (doi:10.1038/378212a0)

Hall, T. M. T., Porter, J. A., Young, K. E., Koonin, E. V., Beachy, P. A. & Leahy, D. J. 1997 Crystal structure of a hedgehog autoprocessing domain: homology between hedgehog and self-splicing proteins. *Cell* **91**, 85–97. (doi:10.1016/S0092-8674(01)80011-8)

Hino, K., Satou, Y., Yagi, K. & Satoh, N. 2003 A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. VI. Genes for Wnt, TGFbeta, Hedgehog and JAK/STAT signaling pathways. *Dev. Genes Evol.* **213**, 264–272. (doi:10.1007/s00427-003-0318-8)

Hwang, J. J. & Stupp, S. I. 1999 Modifying tissue surfaces by liquid crystal formation. *United States Patent*: 6,420,519.

James-Clark, H. 1866 Note on the Infusoria Flagellata and the Spongiae Ciliatae. *Am. J. Sci.* **1**, 113–114.

Jones, D. T. 1999 Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **287**, 797–815. (doi:10.1006/jmbi.1999.2583)

Katti, M. V., Sami-Subbu, R., Ranjekar, P. K. & Gupta, V. S. 2000 Amino acid repeat patterns in protein sequences: their diversity and structural–functional implications. *Protein Sci.* **9**, 1203–1209.

King, N. & Carroll, S. B. 2001 A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution. *Proc. Natl Acad. Sci. USA* **98**, 15 032–15 037. (doi:10.1073/pnas.261477698)

Lang, B. F., O'Kelly, C., Nerad, T., Gray, M. W. & Burger, G. 2002 The closest unicellular relatives of animals. *Curr. Biol.* **12**, 1773–1778. (doi:10.1016/S0960-9822(02)01187-9)

Lee, J. J., von Kessler, D. P., Parks, S. & Beachy, P. A. 1992 Secretion and localized transcription suggest a role in positional signaling for products of the segmentation gene hedgehog. *Cell* **71**, 33–50. (doi:10.1016/0092-8674(92)90264-D)

Mann, R. K. & Beachy, P. A. 2000 Cholesterol modification of proteins. *Biochim. Biophys. Acta* **1529**, 188–202.

Pettitt, M. E., Orme, B. A. A., Blake, J. R. & Leadbeater, B. S. C. 2002 The hydrodynamics of filter feeding in choanoflagellates. *Eur. J. Protistol.* **38**, 313–332. (doi:10.1078/0932-4739-00854)

Philippe, H., Snell, E. A., Bapteste, E., Lopez, L., Holland, P. W. H. & Casane, D. 2004 Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol. Biol. Evol.* **21**, 1740–1752. (doi:10.1093/molbev/msh182)

Porter, J. A., von Kessler, D. P., Ekker, S. C., Young, K. E., Lee, J. J., Moses, K. & Beachy, P. A. 1995 The product of *hedgehog* autoproteolytic cleavage active in local and long-range signalling. *Nature* **374**, 363–366. (doi:10.1038/374363a0)

Shimeld, S. M. 1999 The evolution of the hedgehog gene family in chordates: insights from amphioxus hedgehog. *Dev. Genes Evol.* **209**, 40–47. (doi:10.1007/s004270050225)

Simpson, P. J., Bolam, D. N., Cooper, A., Ciruela, A., Hazlewood, G. P., Gilbert, H. J. & Williamson, M. P. 1999 A family IIb xylan-binding domain has a similar secondary structure to a homologous family IIa cellulose-binding domain but different ligand specificity. *Structure* **7**, 853–864. (doi:10.1016/S0969-2126(99)80108-7)

Snell, E. A., Furlong, R. F. & Holland, P. W. H. 2001 Hsp70 sequences indicate that choanoflagellates are closely related to animals. *Curr. Biol.* **11**, 967–970. (doi:10.1016/S0960-9822(01)00275-5)

Taylor, W. R. 1997 Multiple sequence threading: analysis of alignment quality and stability. *J. Mol. Biol.* **269**, 902–943. (doi:10.1006/jmbi.1997.1008)

Taylor, W. R. & Stoye, J. P. 2004 Consensus structural models for the amino terminal domain of the retrovirus restriction gene *Fv1* and the murine leukaemia virus capsid proteins. *BMC Struct. Biol.* **4**, 1. (doi:10.1186/1472-6807-4-1)

Tomme, P., Warren, R. A. J., Miller Jr, R. C. & Gilkes, N. R. 1995 Cellulose-binding domains: classification and properties. In *Enzymatic degradation of insoluble carbohydrates* (ed. J. N. Saddler & M. H. Penner). American Chemical Society Symposium Series 618, pp. 142–163. Washington, DC: American Chemical Society.

Zardoya, R., Abouheif, E. & Meyer, A. 1996 Evolution and orthology of hedgehog genes. *Trends Genet.* **12**, 496–497. (doi:10.1016/S0168-9525(96)20014-9)

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.